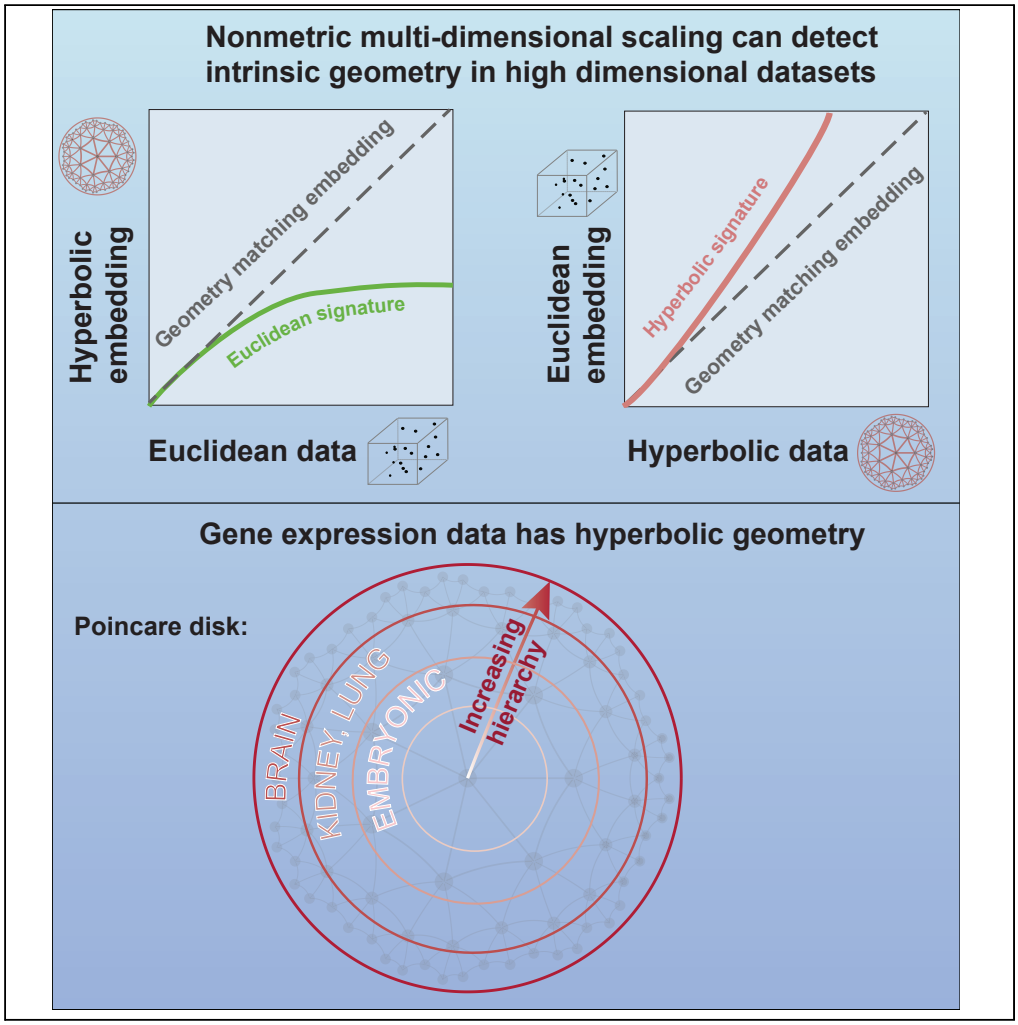
iScience

CellPress
OPEN ACCESS

Article

Hyperbolic geometry of gene expression



Yuansheng Zhou, Tatyana O. Sharpee

sharpee@salk.edu

HIGHLIGHTS

A method to identify underlying low-dimensional geometry in high-dimensional dataset

Gene expression data exhibit local Euclidean and large-scale hyperbolic geometry

The size of the hyperbolic map is larger in differentiated and brain cells

Taking into account hyperbolic geometry yields improved visualization of data

Zhou & Sharpee, iScience 24, 102225
March 19, 2021
https://doi.org/10.1016/j.isci.2021.102225

iScience

CellPress
OPEN ACCESS

Article

Hyperbolic geometry of gene expression

Yuansheng Zhou[1,2] and Tatyana O. Sharpee[1,3,4,*]

SUMMARY

Patterns of gene expressions play a key role in determining cell state. Although correlations in gene expressions have been well documented, most of the current methods treat them as independent variables. One way to take into account gene correlations is to find a low-dimensional curved geometry that describes variation in the data. Here we develop such a method and find that gene expression across multiple cell types exhibits a low-dimensional hyperbolic structure. When more genes are taken into account, hyperbolic effects become stronger but representation remains low dimensional. The size of the hyperbolic map, which indicates the hierarchical depth of the data, was the largest for human cells, the smallest for mouse embryonic cells, and intermediate in differentiated cells from different mouse organs. We also describe how hyperbolic metric can be incorporated into the t-SNE method to improve visualizations compared with leading methods.

INTRODUCTION

One of the great challenges of modern biology is to understand how the genotype of an organism impacts its phenotype, such as disease risk. The difficulty of this problem stems from the complexity of this relationship where thousands of genes can affect a phenotype of interest through nonlinear interactions (The Wellcome Trust Case Control Consortium, 2007; Manolio et al., 2009; Yang et al., 2010). In the past 15 years, genome-wide association studies have demonstrated that a range of traits, including those that are related to metabolic and mental health disorders, are potentially linked to thousands of genes, with each gene explaining only a small fraction of the expected heritability (Manolio et al., 2009). At the same time, correlations between genes are widespread (Novembre et al., 2008). These observations raise the possibility that genetic variation and their expression can be described by a low-dimensional geometry. Identifying this geometry would make it easier to find relevant gene combinations and how they impact a given trait.

Traditional approaches to finding low-dimensional spaces, such as the principal-component analysis (PCA), assume that the space is "flat" (i.e., has zero curvature) and evaluate distances between points according to Euclidean metric. Recently, hyperbolic spaces have attracted a lot of attention both for the analysis of biological data (Zhou et al., 2018; Klimovskaia et al., 2020; Ding and Regev, 2019) and in computer science (Wilson et al., 2014; Nickel and Kiela, 2017; Walter and Ritter, 2002; Shavitt and Tankel, 2008; Cvetkovski and Crovella, 2017; Ovinnikov, 2019; Ganea et al., 2018). The reason for this interest is that hyperbolic metric approximates the exponential expansion of possible states of the system described by a hierarchical tree-like process (Krioukov et al., 2010). Hierarchical representations, such as phylogenetic trees and clustering clades have long been used to characterize differences between cells (Eisen et al., 1998), proteins (Manning et al., 2002), the activity of metabolic networks within cells (Ravasz et al., 2002; Dunkel et al., 2014), and human brain functional networks (Meunier et al., 2009). This suggests that hyperbolic metric should be considered as one of the possibilities when searching for the low-dimensional geometry in biological data. At the same time, any hyperbolic geometry (which has negative curvature) can locally be approximated using Euclidean geometry (which has zero curvature). Therefore, in this work we focus on comparing the signatures of Euclidean and hyperbolic geometry. For completeness, we also include results from the spherical geometry that has positive curvature and represents the remaining of three possible geometries with constant curvature.

In this work we pursue two goals. The first goal is to develop a quantitative test for distinguishing the curvature of the underlying low-dimensional geometry. We show that this can be achieved by performing non-metric multi-dimensional scaling using both Euclidean and hyperbolic metric and comparing the results. Our second goal is to develop visualization tools for data that exhibit a low-dimensional hyperbolic geometry. Many of the current state-of-the-art visualization tools, such as k-means clustering (MacQueen, 1967), local linear embedding (Roweis and Saul, 2000), t-distributed Stochastic Neighbor Embedding (t-SNE)

[1]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[2]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA

[3]Department of Physics, University of California San Diego, La Jolla, CA 92093, USA

[4]Lead contact

*Correspondence: sharpee@salk.edu

https://doi.org/10.1016/j.isci.2021.102225

(Maaten and Hinton, 2008; Zhou and Sharpee, 2018), and Uniform Manifold Approximation and Projection (UMAP) (Becht et al., 2019), all use Euclidean metric. We propose a method for incorporating hyperbolic metric into the t-SNE method and show that this leads to improved visualization across a range of datasets.

To demonstrate the utility of both the diagnostic method and the hyperbolic t-SNE (h-SNE), we apply these methods to a range of gene expression datasets from mouse and human. These datasets uniformly show that gene expression data across different cell types exhibit a low-dimensional hyperbolic geometry. The curvature of this space, which is related to the branching ratio of the corresponding tree-like process, was systematically higher in differentiated cell types compared with embryonic cells and took even larger values for brain cells. These results demonstrate that gene expression data can be effectively described using a small number of coordinates under hyperbolic metric. Visualizations using hyperbolic metric consistently showed more accurate representations, both in terms of local and large-scale structure, including more consistent estimates of developmental states in datasets where pseudo-time trajectories could be constructed (Klimovskaia et al., 2020).

RESULTS

Non-metric MDS outperforms metric MDS in geometry detection

Multi-dimensional scaling (MDS) has been widely used to embed a set of data points into a geometric space in a way that attempts to best preserve the distances between points in the original space. Metric MDS tries to make the embedding distances proportional to the input distances, whereas non-metric MDS only preserves the ordinal values, allowing a monotonic nonlinear transformation between the distances. Both metric and non-metric MDS in high-dimensional Euclidean space have been well studied during the past few decades. However, the MDS in the hyperbolic space has not been fully developed yet. Several metric MDS algorithms have been proposed recently for embedding data into hyperbolic space, offering advantages over Euclidean visualizations in terms of distance preservation, space capacity, trajectory inference, and unseen data prediction (Sala et al., 2018; Klimovskaia et al., 2020; Wilson et al., 2014; Nickel and Kiela, 2017; Walter and Ritter, 2002; Shavitt and Tankel, 2008; Cvetkovski and Crovella, 2017; Ovinnikov, 2019; Ganea et al., 2018; Ding and Regev, 2019), etc. However, we find that metric MDS does not correctly distinguish between Euclidean and hyperbolic geometry of input data, but non-metric MDS does (Figure 1). The reason for this is that non-metric MDS matches the ranking order instead of exact values of the data distances. The resulting nonlinear distortions in embedding distances can be used as indicators for a geometry mismatch between data and embedding points. When using non-metric MDS, we illustrate that as soon as there is a mismatch between native and embedding geometry, a nonlinear distortion appears in the scatterplots of embedding distances versus input data distances (Figures 1A and 1B). These scatterplots are known as Shepard diagrams (Shepard, 1980). When Euclidean data are embedded into a hyperbolic space, the Shepard diagram has negative convexity (Figure 1A). When hyperbolic data are embedded into Euclidean space, the Shepard diagram has a positive convexity (Figure 1B). Thus the convexity of the Shepard diagram can indicate the difference in geometric properties between the embedding and native spaces, and in particular could indicate the difference in curvature of geometry. When using the metric MDS, the Shepard diagram shows increased spread (Figure 1D) but does not yield a nonlinear relationship upon embedding Euclidean data to hyperbolic space (Figure 1C). The reason is that Euclidean distances can be fully embedded into the faster-expanding hyperbolic space masking the distortion of distances, and this does not happen in the non-metric MDS (Shepard, 1980). In what follows we apply non-metric MDS to synthetic and several real gene expression datasets to detect their hidden geometry, and we refer to non-metric MDS as simply MDS for brevity.

Synthetic geometric data

When cells are characterized according to the expression of thousands of genes, the number of genes represents the nominal dimension of the representation space. However, the real dimension of the gene expression space might be much lower. Furthermore, the true geometry of the hidden space is not necessarily Euclidean. Therefore, in this section we analyze the signatures of low-dimensional geometry of constant curvature (either Euclidean, spherical, or hyperbolic) in the situation where each data point is described with respect to large number of variables. In the synthetic examples below, the points are first sampled from a low-dimensional geometry and then embedded into a high-dimensional Euclidean space. This step is included to mimic analysis of experimental data, where each data point is evaluated according to a large number of measurements. After this, the data points are embedded into spaces of different curvatures to determine indicators through which the properties of the original low-dimensional space can become apparent. In the examples below, we focus primarily on hyperbolic and Euclidean geometries,

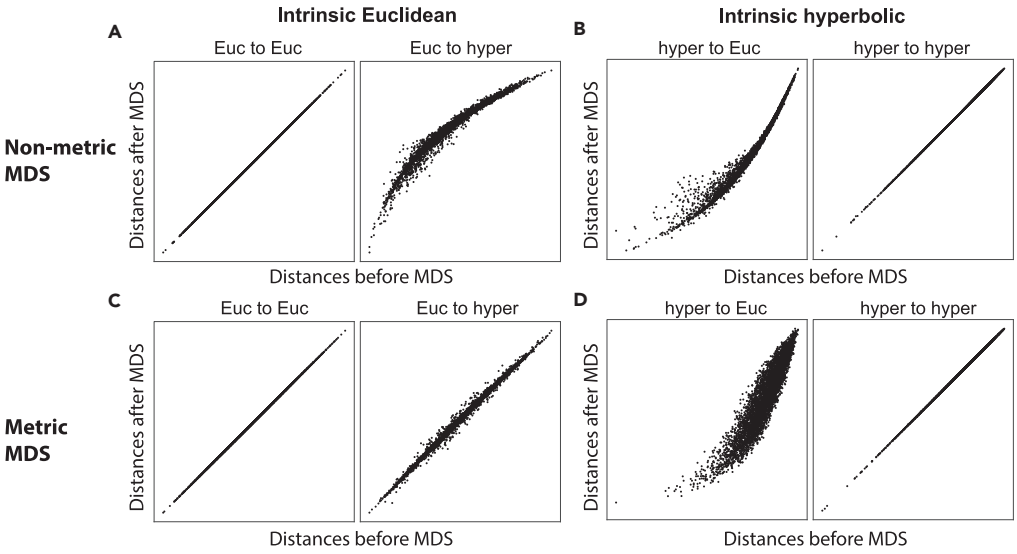


Figure 1. Shepard diagrams for metric MDS and non-metric MDS applied to synthetic geometric data with either Euclidean or hyperbolic native geometry

(A–D) These simulations were produced by (1) randomly sampling 100 points in 5D Euclidean (A and C) and hyperbolic space (B and D), (2) computing geometric distances using the corresponding distance metrics, and (3) using either non-metric MDS (A and B) or metric MDS (C and D) to embed the points to 5D Euclidean and hyperbolic space. The radius of the hyperbolic space (measured in units of inverse curvature) was 3.0 for both the sampling and embedding spaces. The convexity of Shepard diagrams reflects the difference in geometry between the embedding and native spaces; it is positive when hyperbolic data are embedded in Euclidean space and negative when Euclidean data are embedded into the hyperbolic space. These differences are less distinct in the case of metric MDS (bottom row).

because hyperbolic geometry describes hierarchically organized data, whereas Euclidean metric is often the only feasible geometric metric for computing distances of high-dimensional vectors. Comparison with the results for spherical spaces is provided in Figure S1.

First, we analyze the case where data have a 5D Euclidean underlying geometry. To simulate this case we randomly sample 100 points from a 5D Euclidean space and use Euclidean MDS (EMDS) to embed the points to 5D, 10D, 50D, and 100D space, respectively (Figure 2A, left). This step emulates the representation of real data where each data point is described by a large number of measurements (e.g., transcriptome) according to which each cell is characterized, and the distances between points are measured according to a Euclidean metric. The embeddings with different number of dimensions correspond to cases where measurements are taken with respect to different number of genes. As expected, the distances of synthetic 5D Euclidean points can be preserved without distortion when embedding data to Euclidean spaces of higher dimensions. This is evidenced by the linearity of Shepard diagrams in the left column of Figure 2A. Next, we apply EMDS (Figure 2A, middle) and hyperbolic MDS (HMDS) (Figure 2A, right) to the points in the Euclidean representation space, as we did in Figures 1A and 1B. As one can see in Figure 2A, Euclidean embeddings of these data do not generate distortions in the Shepard diagrams, but hyperbolic embeddings yield Shepard diagrams with negative convexity that is largely independent of embedding dimensions. This indicates that the data have an underlying Euclidean geometry.

To quantitatively characterize Shepard diagrams we fit them using:

$$y = a \cdot (x - x_0)^{\kappa+1}, \qquad \text{(Equation 1)}$$

where x and y represent distances of points before and after embedding, respectively; parameter $x_0 = \min(x) - \varepsilon$ is the distance offset representing the difference caused by noise from biological variations or experimental measurements, with a small ε introduced to avoid zero input values in the fitting. The parameter κ is key because it characterizes the convexity of Shepard diagrams. The zero $\kappa = 0$ indicates pure linearity and an exact match between the model and data geometries, whereas $\kappa \neq 0$ indicates convexity and a mismatch between the two geometries. So the sign of κ can indicate the difference in curvature between two spaces. In the current examples, $\kappa = 0$ in EMDS and $\kappa = -0.5$ in HMDS embedding

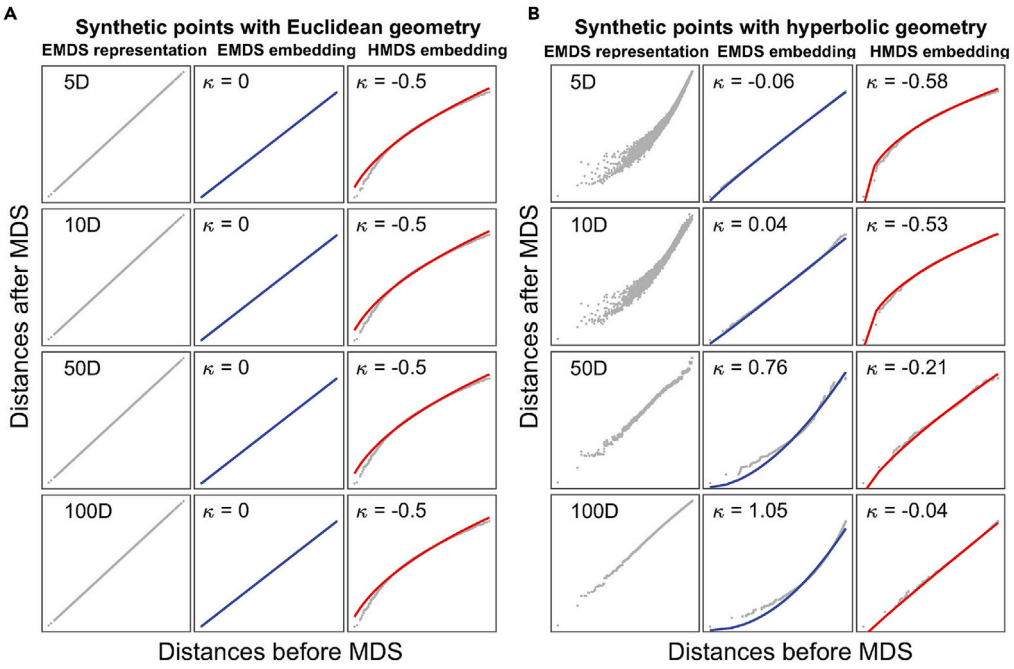


Figure 2. Illustration of the diagnostic approach based on MDS for geometry detection on synthetic data

(A) Randomly sampled 100 points from 5D Euclidean space are embedded into 5, 10, 50, and 100 dimensional Euclidean spaces (left), followed by subsequent embeddings into 5D Euclidean space (middle) or hyperbolic space (right). The solid lines represent the fits using Equation (1), the Shepard diagram curvatures κ are shown at the top of each panel.

(B) Same analysis for 100 sampled points from a 5D hyperbolic space with R_{data} = 3.0. The hyperbolic radii used in HMDS are R_{model} = 3.0 in both (A) and (B).

with R_{model} = 3, with no changes as the representation dimension (Figure 2A). These values describe the signatures of data that have intrinsic Euclidean geometry (Figure 1A).

The situation is qualitatively different for the case where the data have a hidden hyperbolic geometry, cf. Figure 2B. Here we sample 100 points from a 5D hyperbolic space with R_{data} = 3.0. The initial embedding of these points into a low-dimensional Euclidean space produces distortions, indicating that using Euclidean metric to evaluate hyperbolic distances between points will not be accurate when using the same dimension for the embedding space. However, Euclidean embeddings into larger dimensional spaces can produce accurate distance representations. For example, in the left column of Figure 2B, accurate distance representation is obtained starting with ~50 embedding dimensions. The reason for this is that in a large-dimensional Euclidean space points could be distributed along hyperbolic manifolds, approximating the true hyperbolic metric. We next apply MDS to examine the geometry of the representations by embedding them into a low-dimensional Euclidean (middle column) or hyperbolic space (right column). With the increase of representation dimension, the convexity parameter κ of the Shepard diagram increases from approximately zero value ($\kappa = -0.06$) to $\kappa = 1.05$ in EMDS and increases from $\kappa = -0.58$ to an approximately zero value ($\kappa = -0.04$) in HMDS (Figure 2B). These signatures (Figure 1B) indicate that hyperbolic property is more fully preserved when points are characterized with respect to more dimensions. These analyses of synthetic data illustrate how a combination of EMDS and HMDS can be used to elucidate the intrinsic geometry starting with the initial Euclidean representation. This method can also be used to detect spherical geometry, which has positive curvature. Synthetic results show that spherical geometry has opposite property as hyperbolic geometry: spherical is to Euclidean looks like what Euclidean is to hyperbolic in Shepard diagram (Figure S1).

Geometry of gene expression data

We now apply this method to analyze the intrinsic geometry of gene expression data. We first analyze a discrete gene expression data from Lukk et al. (2010). In the article, they integrated microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states, and cell lines, which has a complex global structure. They constructed a global gene expression map by performing

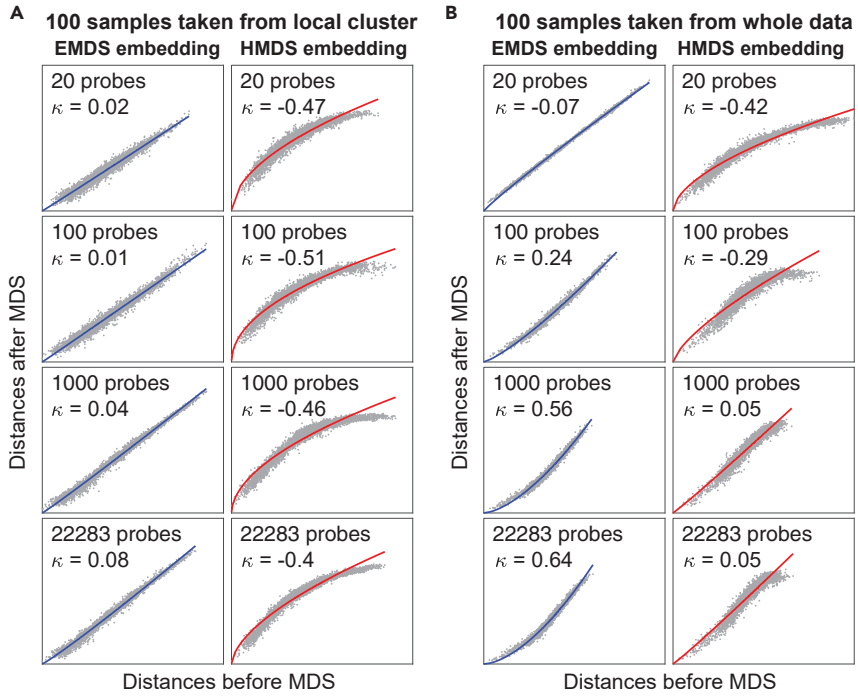


Figure 3. Human gene expression has locally Euclidean and globally hyperbolic hidden geometry

(A) MDS embedding results for samples taken from a single k-means cluster, with distances evaluated by Euclidean metric with respect to increasing number of probes. Left and right columns show results of embeddings into 5D Euclidean and hyperbolic space, respectively.

(B) Same analysis for samples taken randomly from the whole data. The hyperbolic space used in HMDS had radius $R_{model} = 2.6$ in both (A) and (B).

PCA and found that the first two principal axes described variation in biological variables corresponding to hematopoietic and malignancy properties. However, the presence and properties of the underlying low-dimensional geometry and how the samples are organized in the space remain to be investigated. Several previous studies showed that gene expression was stochastic both at the single cell level and the population level (Elowitz et al., 2002; Oleksiak et al., 2002; Raj and van Oudenaarden, 2008), and the expression profiles of samples within the same cluster were dominated by intrinsic noise (Elowitz et al., 2002). This would imply either Euclidean geometry, at least locally, or a lack of geometric structure altogether. On a global scale, biological systems usually show a hierarchical structure, which would imply hyperbolic geometry (Ravasz et al., 2002; Meunier et al., 2009). Therefore, we separately probe the geometry of gene expression data at the local and global scales. To probe local geometry we apply k-means ($k = 50$) method to cluster the whole data and select 100 samples from a single cluster randomly. Similarly to Figure 2, we use increasing subsets of genes (from 20 probes to all the 22,283 probes) to represent samples and then perform EMDS and HMDS embeddings ($R_{model} = 2.6$) for geometry detection (Figure 3A). Increasing the number of probes with respect to which samples were characterized corresponds to increasing the dimensionality of the initial Euclidean embedding as in Figure 2. We find that this does not significantly change the convexity of the Shepard diagram in both EMDS and HMDS ($\kappa \approx 0$ in EMDS and $\kappa \leq -0.4$ in HMDS). These results match the fitting in Figure 2A and indicate that the samples taken from the same cluster have Euclidean structure, even when all the probes are used (Figure 3A). Additional analyses show that the Euclidean structure is indeed caused by the stochastic Gaussian expressions of genes among the samples within a cluster (Figure S2).

Variations in gene expression across samples taken from different clusters, which represent different cell types, tissues, and disease states, show more complicated distributions (Figure S2) and have attracted a great deal of attention (Aguet et al., 2017). To study the geometric structure of expression space globally, we selected 100 samples randomly from the whole population instead of local clusters and performed the same embeddings as in Figure 3A. Surprisingly we find that, as the number of probes increases, the

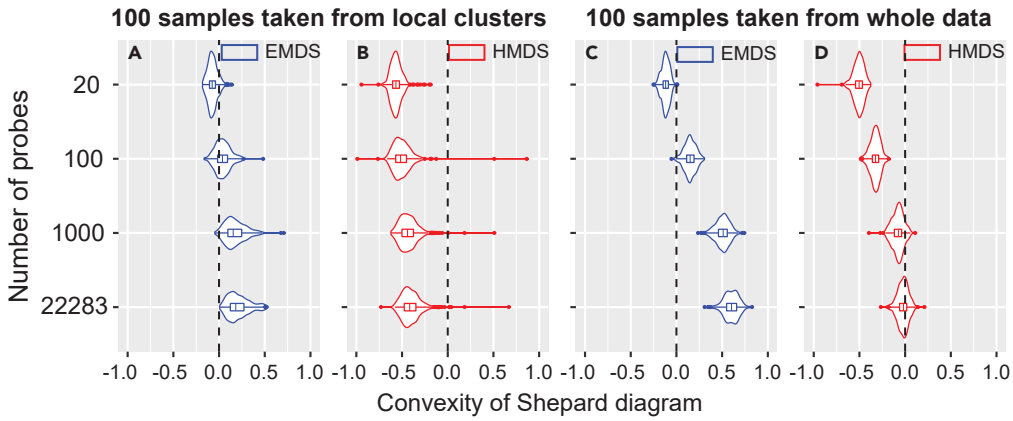


Figure 4. Statistics of convexity parameter of Shepard diagrams across human gene expression data

(A and B) Violin plots show the convexity statistics in 5D EMDS (A) and 5D HMDS (B) across 300 repeated samplings from different k-means clusters, as a function of the number of probes. 100 data points are taken in each sampling.

(C and D) Same analysis for samples taken with replacement from the whole data. The black dashed lines show $\kappa = 0$ and signify Euclidean geometry in EMDS (A and C) or hyperbolic geometry with $R_{data} = R_{model} = 2.6$ in 5D HMDS (B and D). In each plot, the width of the shape shows the probability density of different values; the central line, the left edge and right edge of the box within the shape represent the median, the 75th, and the 25th percentiles respectively. The line within the shape extends to the most extreme non-outlier points; the outliers are represented by dots.

convexity of the Shepard diagram increases from being approximately zero $\kappa = -0.07$ to $\kappa = 0.64$ in EMDS and from $\kappa = -0.42$ to $\kappa = 0.05$ in HMDS (Figure 3B). These fitting results match the signatures expected for hyperbolic geometry in Figure 2B. It shows that the gene expression space has hyperbolic structure that becomes increasingly more apparent upon including a moderately large number of genes (>1,000 probes) in the measurements.

To test the robustness of this conclusion and make full use of the whole data, we repeat the sampling process 300 times both for the local sampling where samples are taken from different single clusters and for the global sampling where samples are broadly taken from the whole data. The samples are taken with replacement. As expected, for samples taken from local clusters, the median values of convexity $\kappa \approx 0$ in EMDS and $\kappa < 0$ in HMDS (Figures 4A and 4B) even when all genes are used. These measurements indicate Euclidean structure. For samples taken across the whole population, with increasing number of probes, the median of κ increases to be positive in EMDS and close to zero in HMDS (Figures 4C and 4D); these signatures indicate that samples across population have hyperbolic structure when represented by a moderately large number of genes ($\geq$1,000 probes).

The size of the hyperbolic gene expression map varies systematically across cell types

The HMDS method can also be used to estimate the curvature of the underlying low-dimensional space or equivalently the size of the hyperbolic map measured in units of inverse curvature (Figure S3). The hyperbolic radius R_{data} can be used as an indication of the hierarchical depth of the corresponding tree structure (Krioukov et al., 2010). Above we have shown that global human gene expression data can be embedded without distortion to 5D hyperbolic space with $R_{model} = 2.6$. This value was obtained by systematically screening across different R values to find those best matching R_{data} as indicated by the zero convexity parameter κ obtained by fitting the corresponding Shepard diagram. In Figure 5A we show that the convexity κ decreases with R_{model} and crosses 0 at $R_{model} \approx 2.6$. Therefore, we conclude that $R_{data} = 2.6$ in 5D hyperbolic representation. Next, we examine several other gene expression datasets and determine R_{data} for them. Han et al. (2018) performed Microwell-seq of cells from multiple mouse organs and generated mouse cell atlas map using the t-SNE method. Here we re-analyze these data to find if they have a nonlinear low-dimensional structure. The microwell-seq data in this dataset are much sparser than the microarray data. Therefore, we first check whether changes in sparseness of measurement data could affect the geometry detection and its parameters. To this end, we re-analyze synthetic data where at the stage of Euclidean high-dimensional embedding all values are re-set to zero if their values are in the smallest 5%. Even though intermediate embeddings into larger dimensional space have more values that are set to zero, this does

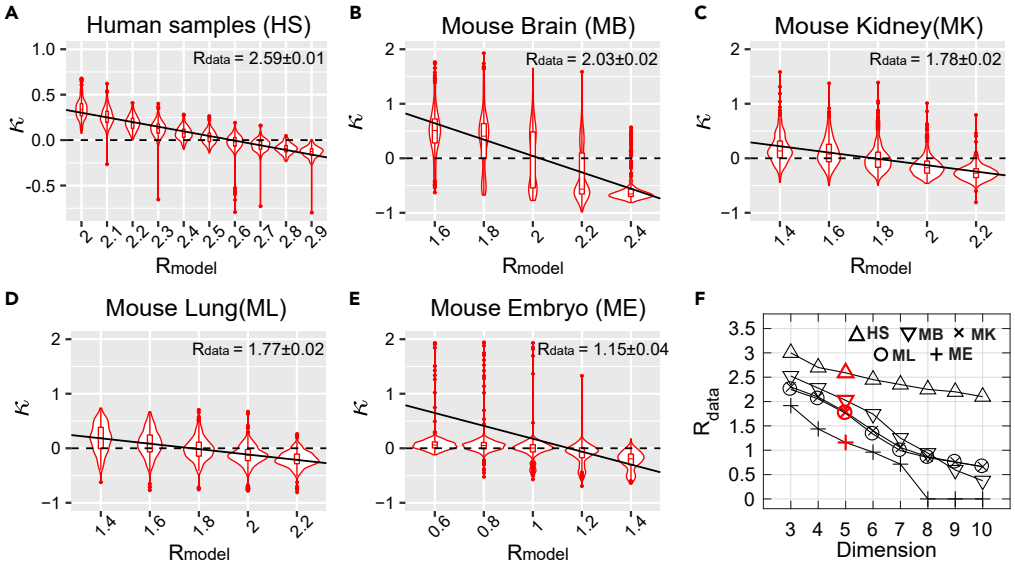


Figure 5. Radius of the hyperbolic space of gene expression varies systematically across cell types

(A–E) The violin plots of convexity of Shepard diagrams κ as a function of the radius of the embedding space R_{model}. (A) Microarray data from human samples in Lukk et al. (2010) dataset. (B–E) Microwell-seq data (Han et al., 2018) for brain cells (B), kidney (C), lung (D), and embryonic stem cells (E). Solid lines show the linear regression; their *y*-intercepts yield estimates of the radius of the hyperbolic map for each dataset (see printed values with ±2 SD from 300 samplings). The HMDS embedding dimension is $D = 5$ in (A–E).

(F) Dependence of the radius of the hyperbolic space as a function of the embedding dimension for datasets in (A–E) on human cells (HS), mouse brain (MB), mouse kidney (MK), mouse lung (ML), and mouse embryonic cells (ME). $R_{data} = 0$ ($D \geq 8$ in ME) means that κ remains negative regardless of R_{model}.

not change the estimated convexity values for low-dimensional embeddings using either Euclidean or hyperbolic metric, and the tests correctly identify the presence of a low-dimensional hyperbolic geometry (Figure S4).

With these checks at hand, we proceeded to analyze the microwell-seq data from different mouse organs. Following previous studies (Macosko et al., 2015), the data were pre-processed using the Seurat algorithm (Butler et al., 2018) and projected onto top 50 principal components (see transparent methods). Next, we applied 5D HMDS to the processed data from four of the mouse organs – brain, kidney, lung, and embryonic stem cells. We find that all these data have an underlying hyperbolic structure (Figures 5B–5E). It is worth noting that the hyperbolic radius necessary to describe these data is smaller than that for human samples. Among the four mouse cell types, the largest radius is found for the mouse brain cells with $R_{brain} = 2.03 \pm 0.02$ (Figure 5B), followed by mouse kidney and lung that have similar radii $R_{data} = 1.78 \pm 0.02$ and 1.77 ± 0.02, respectively (Figures 5B and 5C). Finally, the smallest radius is observed for mouse embryonic stem cells with $R_{data} = 1.15 \pm 0.04$ (Figure 5E). Because hyperbolic radius indicates the depth of the underlying hierarchical tree, these findings indicate an interesting progression in complexity with embryonic cells exhibiting the smallest degree of hierarchical organization and brain cells exhibiting the largest degree.

We note that the HMDS methods produce estimates of the hyperbolic radius that depend on the embedding dimension *D*. This happens because the density of points increases exponentially with exponent $(D-1)$ *R* according to Equation (S3). Results in Figures 5A–5E are obtained for a 5D hyperbolic space. In panel F, we show how the estimates of R_{data} decrease with embedding dimension in different datasets (Figure 5F). Importantly, the relative differences in R_{data} across cell types are maintained across a range of different embedding dimensions. The hyperbolic maps continue to have the smallest radius for mouse embryonic cells, larger values for mouse differentiated cells, and yet larger values for mouse brain and human cells. We also find that the minimal embedding dimension for all of these datasets is $D = 3$, and that smaller dimension fails to properly embed the data.

We also tested the robustness of the HMDS method to noise in the data. Toward that goal, we add varying amounts of the multiplicative Gaussian noise to the Lukk et al. data (Lukk et al., 2010) and fit the resulting Shepard diagrams. The fits produce stable convexity estimates for Shepard's diagrams over a broad range of noise values (Figures S5A–S5E). This robustness is observed up to very large noise values with $\varepsilon = 0.5$ when noise completely destroys the data structure (Figure S5F). The reason for this robustness is that noise does not systematically shift the shape of the Shepard diagram, yielding the same fitting exponent under varying noise amounts.

Hyperbolic low-dimensional visualization of gene expression data

Although MDS embedding can be used to detect intrinsic geometry, it is not ideal for low-dimensional visualization. One of the primary reasons common to all MDS-based algorithms is that they are not designed to attract similar points together like t-SNE. Consequently, MDS-based methods achieve poor clustering results. These limitations were solved by nonlinear methods like t-SNE and UMAP, which, however, are only performed in the Euclidean space. As a result, existing visualization methods may cause distortion of global structure in the data that has a global hyperbolic structure. Here we aim to adapt the t-SNE algorithm to work in hyperbolic space. To achieve this we use hyperbolic metric to evaluate global distances in the data while keeping the local clustering aspects of the algorithm. The standard t-SNE method effectively discards large distance information between distant points. We recently proposed a variant of t-SNE, which aims to preserve global Euclidean structure in the data, which was called global t-SNE (g-SNE) (Zhou and Sharpee, 2018). The g-SNE method works by adding to the similarity distance measures present in the t-SNE another term that focuses on large Euclidean distances (see transparent methods). When applied to Lukk et al. data (Lukk et al., 2010), g-SNE preserves data distances very well (Figure 6A, $R = 0.848$). Despite the high quality of embedding, g-SNE cannot reveal the hierarchical structure of data, which is only visible in hyperbolic embedding. Therefore, considering that human gene expression space is locally Euclidean and globally hyperbolic, we develop a hyperbolic t-SNE (h-SNE) method that applies hyperbolic metric to global similarities as defined in g-SNE (Zhou and Sharpee, 2018) while still using Euclidean metric for original local similarities. We find that h-SNE gives similar embedding accuracy as g-SNE, both of which largely outperform PCA and UMAP, with $R = 0.841$ for h-SNE compared with $R = 0.744$ for PCA and $R = 0.627$ for UMAP (Figure 6A). The distance correlation of Shepard diagram generally quantifies the quality of embedding with respect to large distances, i.e., the global inter-class structure preservation. To measure the local structure preservation, we use the silhouette score, which measures the quality of clustering (Rousseeuw, 1987). Here we find that h-SNE achieves higher silhouette score than g-SNE and significantly higher score than other algorithms (Figure 6B).

These quantitative improvements by h-SNE are also reflected in the improved local and global visualizations that the method provides. For local visualization, the clusters identified by h-SNE are well separated with respect to 15 different tissues and disease types (Figure S6). By comparison, the PCA representation does not separate the 15 clusters very well, mixing nervous system neoplasm cells (cyan) with the breast cancer cells (magenta) (Figure 6C). The non-neoplastic cell line (yellow) is also not separated in the PCA representation from the solid tissue neoplasm cell line (green) (Figure 6C, see all the 15 labels in Figure S6). The UMAP methods separate clusters better but generate too many disconnected components that are difficult to be matched to sample labels (Figure S6). In terms of global properties, the h-SNE visualization generates a clearer global hierarchical organization of clusters, which is not attainable in g-SNE embedding: cells from nervous system neoplasm, breast cancer, non-neoplastic cell line, and solid tissue neoplasm cell line are sequentially positioned at different branches in the disk (Figure 6C); in addition, the two principal hematopoietic and malignancy axes can be clearly identified in h-SNE, but not in UMAP (Figure S7). Finally, it is particularly interesting to note the differences in hierarchical positioning that are assigned to breast cancer cells (magenta). Many of these cells occupy points with smaller radii. Positions that are closer to the center of the hyperbolic space typically correspond to more de-differentiated cells, as we have already seen in the comparison between mouse embryonic cells and differentiated cells. Thus, the more central positions assigned to breast cancer cells are consistent with observations of them being close to de-differentiated cells (Friedmann-Morvinski and Verma, 2014).

The quality of h-SNE visualization is also illustrated by the topography with respect to gradient expressions of three marker genes: NCAM1 (Deborde et al., 2016) for nervous system neoplasm, ASPN (Castellana et al., 2012) for breast cancer, and PLOD2 (Song et al., 2017) for non-neoplastic cell line. These marker genes are highly expressed in distinct but continuous branches in h-SNE; by comparison, the expression patterns of these three genes are more difficult to organize in g-SNE, to cluster in UMAP, or to separate in PCA (Figure 6D).

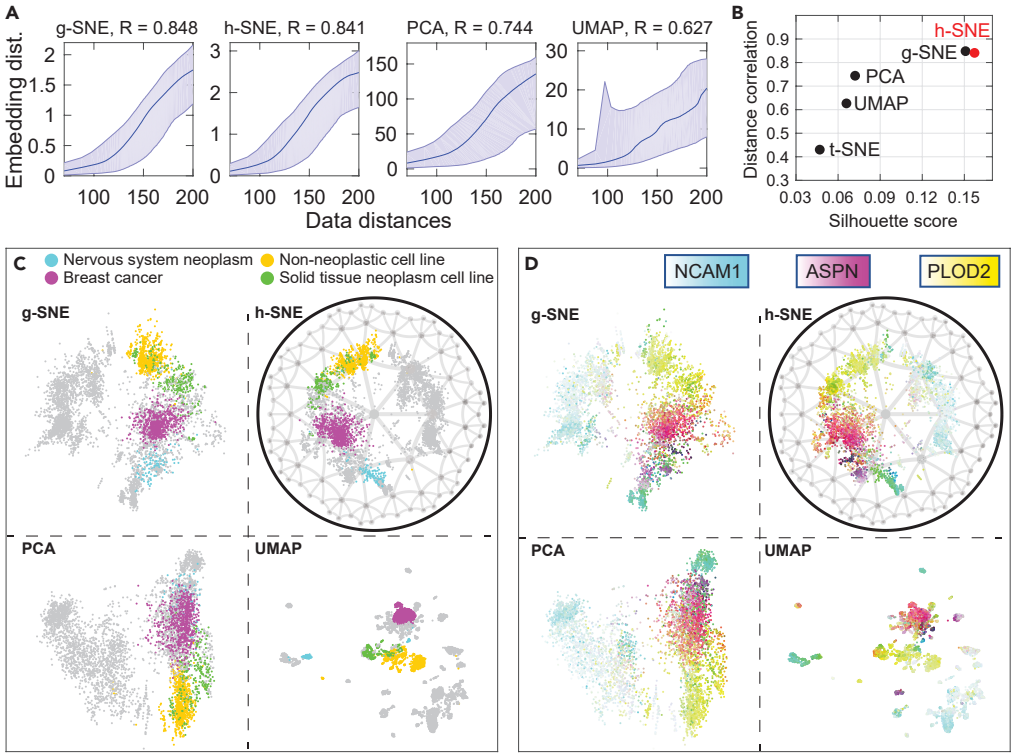


Figure 6. Comparison of two-dimensional visualizations of human expression data (Lukk et al., 2010) using g-SNE, h-SNE, PCA, and UMAP

(A) Shepard diagrams of the four different mappings. The Pearson correlation coefficients of pairwise distances plots are shown at the top of each panel. The shaded regions represent the 95% range of the points at each of the binned data distance intervals (20 bins), the lines in the middle represent the medium.

(B) Quantification of local and global structure preservation using Pearson correlation coefficient of Shepard diagram (*y* axis) and Silhouette score (*x* axis), respectively, for five algorithms (including t-SNE). The Silhouette score is defined as the geometric mean of the three scores obtained by using six hematopoietic labels, four malignancy labels, and fifteen subtype labels, respectively (see transparent methods).

(C) In h-SNE, the data points are visualized within a 2D Poincaré disk with order-7 triangular tiling, which represents a compressed version of a hyperbolic space. In g-SNE, PCA, and UMAP, the points are visualized in 2D Euclidean plane. Four cell types are highlighted with color: nervous system neoplasm (cyan), breast cancer (magenta), solid tissue neoplasm cell line (green), and non-neoplastic cell line (yellow). The rest of the data points are shown in gray to avoid confusion between multiple colors; see Figure S6 for colors across all cell types.

(D) The embedding samples are colored using subtractive CMY color mode according to normalized expressions of three marker genes NCAM1 (nervous system neoplasm), ASPN (breast cancer), and PLOD2 (non-neoplastic cell line).

See also Figure S6.

In addition to visualizing discrete data, hyperbolic embedding is especially useful in representing temporally continuous data and predicting lineage information. Klimovskaia et al. (2020) developed Poincaré map method to visualize hierarchies in single-cell data. This method used similar idea as t-SNE but implemented hyperbolic metric in the representation space. This has led to improvements in the representations of cell trajectories. However, the Poincaré map method, being based on t-SNE, still largely discards large distance information. This problem can be well solved by h-SNE, which is designed to capture global hyperbolic structure. For comparison with the Poincaré map method (Figure 4 in Klimovskaia et al., 2020), we select the mouse hematopoiesis data in Moignard et al. (2015). This dataset consists of cells from different development stages: primitive streak (PS), neural plate (NP), head fold (HF), four somite GFP (Runx1) negative (4SG-), and four somite GFP positive (4SG+). We first apply HMDS method to determine the intrinsic geometry of the data and find that the data space is hyperbolic with R_{data} = 1.72 (Figure S8). Then we apply h-SNE to the data and compare the results with Poincaré map. The h-SNE method produces similar local clustering as in Poincaré map, but it generates very distinct global pattern: the two differentiated branches 4SFG and 4SG extend around the disk with clear division along the angular variable in the h-SNE

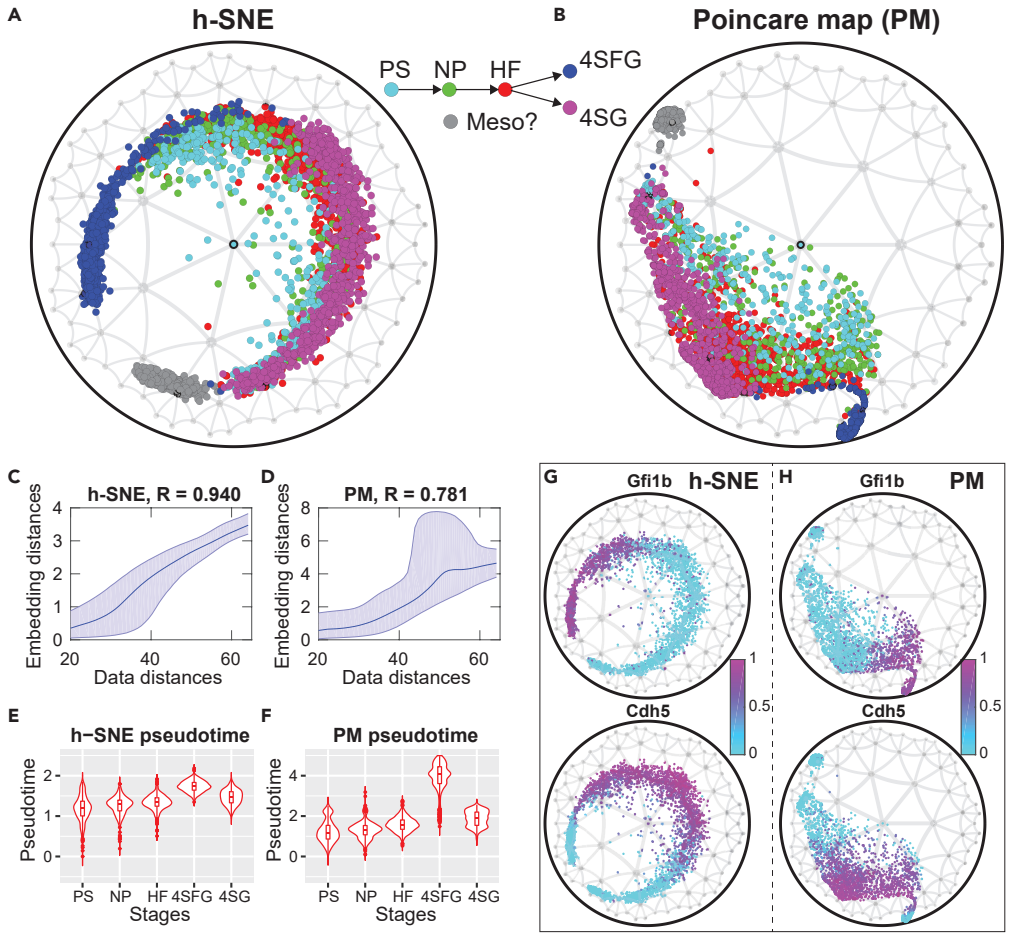


Figure 7. Comparison of hyperbolic embedding of mouse hematopoiesis data using h-SNE and Poincaré map

(A and B) h-SNE and Poincaré mapping of the data after centering the root node in Poincaré disk (see transparent methods). Gray cluster represents potential outliers or "mesodermal" cells (Klimovskaia et al., 2020) (Moignard et al., 2015).

(C and D) Shepard diagram of h-SNE and Poincaré mapping. The data distances are calculated using Euclidean metric in the original data space, whereas the embedding distances are calculated using hyperbolic metric.

(E and F) Predicted pseudo time from h-SNE and Poincaré mapping; the pseudo time is defined as the hyperbolic distances between points and the root node.

(G and H) Normalized gene expressions of two main genes Gfi1b (hemogenic marker) and Cdh5 (endothelial marker) in h-SNE and Poincaré maps.

visualization (Figure 7A). The corresponding pattern is not as clear in Poincaré map (Figure 7B). The Shepard diagrams of the embeddings show that h-SNE preserves data distances much better than Poincaré map, especially the large distances (Figures 7C and 7D). When predicting the pseudo time, the h-SNE method produces a clear pseudo time prediction with a much smaller variance compared with the Poincaré map (c.f. compare the pseudo time in 4SFG stage, Figures 7E and 7F). Finally, as another example, we show the normalized gene expressions of two marker genes Gfi1b (hemogenic marker) and Cdh5 (endothelial marker), finding that these two genes are differentially expressed in different branches in h-SNE (Figure 7G). This separation is not obvious in Poincaré map (Figure 7H). The clear hierarchical organization of cells in h-SNE map may help us better understand the relationships between cells at different stages.

DISCUSSION

In this article we developed a non-metric MDS in hyperbolic space and showed how it can be used to detect the hidden geometry of data starting with an initial Euclidean representation. By applying this method to

several gene expression datasets, we found that gene expression data exhibit Euclidean geometry locally and hyperbolic geometry globally. The radius of the hyperbolic space differed depending on the cell types. The lowest values were observed for embryonic cells and the highest values were observed for brain cells in mouse data. Given that hyperbolic geometry is indicative of hierarchically organized data (Zhou et al., 2018; Krioukov et al., 2010), and the spanned radius represents the depth of the network hierarchy, it is perhaps intuitive that the largest value would be observed for highly differentiated and specialized brain cells and the smallest value for the embryonic cells.

The method that we used to detect the presence of hyperbolic geometry was based on non-metric MDS. One can also use methods from algebraic topology (Zhou et al., 2018; Giusti et al., 2015) for this purpose, as has been recently demonstrated for metabolic networks underlying natural odor mixtures produced by plants and animals (Zhou et al., 2018). The advantage of the topological method is that it is very sensitive to changes in the underlying geometry, including its dimensionality and hyperbolic radius. However, this method is computationally intensive and does not scale well to large datasets. In contrast, the non-metric MDS method is computationally much faster. Therefore, we recommend using it as a first step in determining whether the underlying geometry is hyperbolic or Euclidean. If hyperbolic geometry is detected, then radial position of embedding points can be used to arrange data hierarchically. We have also seen that taking into account hyperbolic geometry produces better low-dimensional visualizations, cf. Figures 6 and 7.

Accurate representation of data across scales is a very active area of research (Wu et al., 2018; Ding et al., 2018; Kobak and Berens, 2018). Special attention is being devoted to developing visualization methods that can not only cluster data in a useful way but also preserve relative positions between clusters (Zhou and Sharpee, 2018; Becht et al., 2019). In particular, preserving global data structure was one of the driving factors for the UMAP method (Becht et al., 2019). Knowing the underlying geometry helps to position clusters appropriately and robustly map them across different runs in a visualization method. For example, the t-SNE method produces random positions of the clusters across different runs of the algorithm (Wattenberg et al., 2016). This problem can in part be alleviated by additional constraints on large distances (Zhou and Sharpee, 2018). Here we find that using a combination of a hyperbolic metric for large distances and Euclidean metric for local distances offers strong improvements in this respect. It also outperforms the recent Poincaré map method that implements hyperbolic metric only for local distances (Klimovskaia et al., 2020). We notice that although h-SNE is best fit for hyperbolic data, it performs similarly as g-SNE in accuracy distances preservation. A future direction is to further optimize h-SNE algorithm.

What could be the origin of hyperbolic geometry at the large scale and Euclidean at small scale? First, any curved geometry, including hyperbolic, is locally flat, i.e., Euclidean. The scale at which non-Euclidean effects become important depends on the curvature of the space. From a biological perspective, the Euclidean aspects can arise from intrinsic noise in gene expression (Elowitz et al., 2002; Oleksiak et al., 2002; Raj and van Oudenaarden, 2008). This noise effectively smoothes the underlying hierarchical process that generates the data. We find that hyperbolic effects of human gene expression can be detected by including measurements on as few as ~100 probes. Why do hyperbolic effects require measurements along multiple dimensions? The reason is that hyperbolic geometry is a representation of an underlying hierarchical process, which generates correlations between variables. These correlations become detectable above the noise once a sufficient number of measurements are made. As an example, one can think of leaves in a tree-like network, and how their activity becomes correlated when it is induced by turning on and off branches of the network. Intuitively, these correlations generate the outstanding branches of a hyperbola. We observe that these correlations can be detected by monitoring even a relatively small (~100) number of probes. This makes it possible to construct a global map of genes from partial measurements, and open new ways for combining data from different experiments.

Limitations of the study

The main limitations of the study pertain to computational constraints. Currently, the hyperbolic MDS methods can reliably embed several hundred data points. Therefore, we analyze data by randomly selecting 100 points at a time. The hyperbolic t-SNE (h-SNE) presented here was developed based on the t-SNE version that performs exact embedding without any speed optimization (e.g., Barnes-Hut t-SNE). As a result, it becomes computationally expensive for large datasets and convergence starts to become problematic when the number of cells becomes too large. The current version performs well with less than ~6,000 cells.

Resource availability

Lead contact

Tatyana Sharpee, sharpee@salk.edu.

Materials availability

The data used in this manuscript are all obtained from published papers.

Data and code availability

The data can be obtained from: Lukk et al. data: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-62/. Han et al. data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108097; and Moignard et al. data: https://github.com/facebookresearch/PoincareMaps/blob/master/datasets/Moignard2015.csv. The codes for non-metric hyperbolic MDS are available from: https://github.com/gyrheart/Hyperbolic-MDS.git. The codes for hyperbolic t-SNE are available from https://github.com/gyrheart/Hyperbolic-t-SNE.git.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102225.

ACKNOWLEDGMENTS

This research was supported by an AHA-Allen Initiative in Brain Health and Cognitive Impairment award made jointly through the American Heart Association and the Paul G. Allen Frontiers Group: 19PABH134610000, Dorsett Brown Foundation, Aginsky Fellowship, NSF grant IIS-1724421, NSF Next Generation Networks for Neuroscience Program (Award 2014217), and NIH grants U19NS112959 and P30AG068635.

AUTHOR CONTRIBUTIONS

Both authors participated in the design of this study and writing of the manuscript. Y.Z. analyzed the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 8, 2020
Revised: December 9, 2020
Accepted: February 18, 2021
Published: March 19, 2021

REFERENCES

Aguet, F., Brown, A.A., Castel, S.J., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A.V., Strober, B.J., Zappal, Z., et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using umap. Nat. Biotechnol. *37*, 38.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Castellana, B., Escuin, D., Peiró, G., Garcia-Valdecasas, B., Vázquez, T., Pons, C., Pérez-Olabarria, M., Barnadas, A., and Lerma, E. (2012). Aspn and gjb2 are implicated in the mechanisms of invasion of ductal breast carcinomas. J. Cancer *3*, 175.

Cvetkovski, A., and Crovella, M. (2017). Low-stress data embedding in the hyperbolic plane using multidimensional scaling. Appl. Math. *11*, 5–12.

Deborde, S., Omelchenko, T., Lyubchik, A., Zhou, Y., He, S., McNamara, W.F., Chernichenko, N., Lee, S.-Y., Barajas, F., Chen, C.-H., et al. (2016). Schwann cells induce cancer cell dispersion and invasion. J. Clin. Invest. *126*, 1538–1554.

Ding, J., Condon, A., and Shah, S.P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat. Commun. *9*, 2002.

Ding, J., and Regev, A. (2019). Deep Generative Model Embedding of Single-Cell Rna-Seq Profiles on Hyperspheres and Hyperbolic Spaces. BioRxiv, 853457.

Dunkel, A., Steinhaus, M., Kotthoff, M., Nowak, B., Krautwurst, D., Schieberle, P., and Hofmann, T. (2014). Natures chemical signatures in human olfaction: a foodborne perspective for future biotechnology. Angew. Chem. Int. Ed. *53*, 7124–7143.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U S A *95*, 14863–14868.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Friedmann-Morvinski, D., and Verma, I.M. (2014). Dedifferentiation and reprogramming: origins of cancer stem cells. EMBO Rep. *15*, 244–253.

Ganea, O., Bécigneul, G., and Hofmann, T. (2018). Hyperbolic neural networks. Advances in Neural Information Processing Systems, 5345–5355.

Giusti, C., Pastalkova, E., Curto, C., and Itskov, V. (2015). Clique topology reveals intrinsic geometric structure in neural correlations. Proc. Natl. Acad. Sci. U S A *112*, 13455–13460.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by microwell-seq. Cell *172*, 1091–1107.

Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. Nat. Commun. *11*, 2966.

Kobak, D., and Berens, P. (2018). The Art of Using T-Sne for Single-Cell Transcriptomics. bioRxiv, 453449.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. (2010). Hyperbolic geometry of complex networks. Phys. Rev. E *82*, 036106.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. (2010). A global map of human gene expression. Nat. Biotechnol. *28*, 322.

Maaten, L.v. d., and Hinton, G. (2008). Visualizing data using t-sne. J. Machine Learn. Res. *9*, 2579–2605.

Macosko, E.Z., Anindita Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *116*, 1202–1214.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability *1*, 281–297.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K., and Bullmore, E.T. (2009). Hierarchical modularity in human brain functional networks. Front. Neuroinform. *3*, 37.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E., et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol. *33*, 269–276.

Nickel, M., and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In Advances in Neural Information Processing Systems, pp. 6338–6347.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within europe. Nature *456*, 98–101.

Oleksiak, M.F., Churchill, G.A., and Crawford, D.L. (2002). Variation in gene expression within and among natural populations. Nat. Genet. *32*, 261.

Ovinnikov, I. (2019). Poincar'e wasserstein autoencoder. arXiv, arXiv:1901.01427.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. Science *297*, 1551–1555.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323–2326.

Sala, F., de Sa, C., Gu, A., and Ré, C. (2018). Representation tradeoffs for hyperbolic embeddings. In International Conference on Machine Learning, pp. 4457–4466.

Shavitt, Y., and Tankel, T. (2008). Hyperbolic embedding of internet graph for distance estimation and overlay construction. IEEE/ACM Trans. Networking (Ton) *16*, 25–36.

Shepard, R.N. (1980). Multidimensional scaling, tree-fitting, and clustering. Science *210*, 390–398.

Song, Y., Zheng, S., Wang, J., Long, H., Fang, L., Wang, G., Li, Z., Que, T., Liu, Y., Li, Y., et al. (2017). Hypoxia-induced plod2 promotes proliferation, migration and invasion via pi3k/akt signaling in glioma. Oncotarget *8*, 41947.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661.

Walter, J.A., and Ritter, H. (2002). On interactive visualization of high-dimensional data using the hyperbolic plane. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), pp. 123–132.

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. Distill *1*, e2.

Wilson, R.C., Hancock, E.R., Pekalska, E., and Duin, R.P. (2014). Spherical and hyperbolic embeddings of data. IEEE Trans. Pattern Anal. Mach. Intell. *36*, 2255–2269.

Wu, Y., Tamayo, P., and Zhang, K. (2018). Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. Cell Syst. *7*, 656–666.

Yang, J., Visscher, P.M., and Wray, N.R. (2010). Sporadic cases are the norm for complex disease. Eur. J. Hum. Genet. *18*, 1039–1043.

Zhou, Y., and Sharpee, T. (2018). Using Global T-Sne to Preserve Inter-cluster Data Structure. bioRxiv, 331611.

Zhou, Y., Smith, B.H., and Sharpee, T.O. (2018). Hyperbolic geometry of the olfactory space. Sci. Adv. *4*, eaaq1458.

iScience, Volume *24*

Supplemental information

Hyperbolic geometry of gene expression

Yuansheng Zhou and Tatyana O. Sharpee

1. Transparent Methods

Metric and non-metric MDS in Euclidean model

Assume there sectionre n objects described by a set of measurements, the dissimilarity of the objects can be obtained by the experimental measurements of the objects. For example, the dissimilarities of two cells can be calculated by the Euclidean distances of the gene expression vectors. Metric MDS approximates the geometric distances d_{ij} to the data dissimilarities δ_{ij}, while non-metric MDS approximates a monotonic transformation of dissimilarities of data. The transformed values are known as disparities $\hat{d_{ij}}$. The loss function S in Euclidean embedding was defined as :

$$S = \sqrt{\frac{S^*}{T^*}} \tag{S1}$$

Where $S^* = \Sigma_{i,j}(d_{ij} - \hat{d_{ij}})^2$, $T^* = \Sigma_{i,j} d_{ij}^2$. In non-metric MDS, $\hat{d_{ij}}$ is determined using the greatest convex minorant method in Kruskal's approach Kruskal (1964). In metric MDS, disparities are equal to dissimilarities: $\hat{d_{ij}} = \delta_{ij}$.

1.1. Non-metric MDS in native hyperbolic model

There are many hyperbolic space representations, we will use the native representation with polar coordinates Krioukov et al. (2010) in our hyperbolic MDS. The angular coordinates in the space are the same as in an Euclidean ball, the radius $R_{\text{model}} \in (0, \infty)$ characterizes the hierarchical depth of the structure, measures the degree of hierarchy in data, and determines how points distribute in the space. The distance of two points d_{ij} is calculated as:

$$\cosh(d_{ij}) = \cosh(r_i)\cosh(r_j) - \sinh(r_i)\sinh(r_j)\cos(\Delta\theta_{ij}) \tag{S2}$$

Where r_i and r_j are the radial coordinates of the two points, and $\Delta\theta_{ij}$ is the angle between them. In D-dimensional HMDS, we initialize the embedding process by uniformly sampling points within radius R_{model} in the native hyperbolic model. The points directions are uniformly sampled around the high-dimensional sphere, and the radial coordinate $r \in (0, R_{\text{model}}]$ follows :

$$\rho(r) \sim \sinh^{D-1} r \tag{S3}$$

We note that there can be merits to sample the points uniformly in the angular variables. Although this does not lead to uniform sampling of points along the sphere Koay (2011), this way of sampling can be particularly advantageous in the situation where the angular variable maps onto periodic variables that correspond to cell cycle or other rhytms. We have used this sampling in our previous publication on olfactory signals produced by fruits and plants Zhou et al. (2018) where it matched developmental processes in the fruit.

During the iteration process, we update both angular and radial coordinates according to the gradient descent of the loss function Eq. (S1), and at the same time set R_{model} as the upper bound of the radial coordinates. The reason of setting a bound is that the coordinates in hyperbolic model are polar coordinates which cannot be normalized after each iteration as performed in Euclidean MDS, so without bound the gradient descent of loss functionsection Eq. (S1) would lead to very large r_i and d_{ij} (since d_{ij} is in the denominator) and hence fail to preserve radial coordinates of data. By setting the upper bound for radial coordinates, the HMDS embedding can well preserve the data distances and precisely detect hyperbolic radius of data R_{data} (Fig. S3).

1.2. Fitting of Shepard diagram

The Shepard diagram is linear if the geometry of input data matches the geometry of embedding space, and otherwise nonlinear. In both EMDS and HMDS, we use the power function below to fit the pairwise distances:

$$y = a(x - x_0)^{\kappa+1} \tag{S4}$$

Where $x_0 = \min(x) - \epsilon$ is an offset representing the distance caused by intrinsic noise of data, a small value ϵ is introduced to avoid zero inputs in the fitting. The convexity κ describes the linearity of the fitting. $\kappa = 0$ indicates Euclidean input in EMDS and means $R_{\text{data}} = R_{\text{model}}$ in HMDS. $\kappa > 0$ means the data is more hyperbolic than the model, and vice versa.

1.3. Hyperbolic t-SNE

Given a data set containing N data points described by D dimensional vectors: $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N; \mathbf{x}_i \in \mathbb{R}^D\}$. The t-SNE algorithm Maaten and Hinton (2008) defines the similarity of two points $\mathbf{x}_i, \mathbf{x}_j$ as the joint probability p_{ij}:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}, \tag{S5}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}. \tag{S6}$$

The similarity of two points $\mathbf{y}_i, \mathbf{y}_j$ in embedding space is defined as the joint probability q_{ij}:

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{m\neq n}(1 + \|\mathbf{y}_m - \mathbf{y}_n\|^2)^{-1}}. \tag{S7}$$

The discrepancy between the similarities of data and embedding points is the loss function, which is defined by Kullback-Leibler (KL) divergence of the joint probability p_{ij} and q_{ij} :

$$L = D_{KL}(P\|Q) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right). \tag{S8}$$

Minimizing the loss function L with respect to the embedding coordinates $\mathbf{y}_i$ by gradient descent gives:

$$\frac{\partial L}{\partial \mathbf{y}_i} = 4\sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}. \tag{S9}$$

The original definitions of similarities Eqs. (S5-S7) in t-SNE are sensitive to small pairwise distances among neighboring points but not to large distances between distant points. To preserve large

distances, Zhou et al. Zhou and Sharpee (2018) proposed global t-SNE algorithm that introduced global similarity terms $\hat{p}_{ij}$ and $\hat{q}_{ij}$ which are primarily sensitive to large distance values:

$$\begin{aligned} \hat{p}_{ij} &= \frac{1 + \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sum_{m \neq n} (1 + \|\mathbf{x}_m - \mathbf{x}_n\|^2)} \\ \hat{q}_{ij} &= \frac{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}{\sum_{m \neq n} (1 + \|\mathbf{y}_m - \mathbf{y}_n\|^2)} \end{aligned} \tag{S10}$$

And they defined the global loss function $\hat{L}$ as:

$$\hat{L} = D_{KL}(\hat{P} \| \hat{Q}) = \sum_i \sum_j \hat{p}_{ij} \log \left(\frac{\hat{p}_{ij}}{\hat{q}_{ij}} \right) \tag{S11}$$

The total loss function L_{total} was then defined by combining the two loss functions using a weight parameter λ:

$$L_{\text{total}} = L + \lambda \hat{L} \tag{S12}$$

The gradient of the total loss function L_{total} gives:

$$\frac{\partial L_{\text{total}}}{\partial \mathbf{y}_i} = 4 \sum_j [(p_{ij} - q_{ij}) - \lambda(\hat{p}_{ij} - \hat{q}_{ij})] \cdot (\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}, \tag{S13}$$

where the weight λ of the global loss function controls the balance between the local clustering and global organization of the data. Large λ values lead to more robust global distribution of clusters, but less clear classifications. Small λ moves back to approximate the traditional t-SNE, and will be exactly the same when $\lambda = 0$. In hyperbolic t-SNE, we still use native representation parametrized by R_{model} as in HMDS, but here R_{model} is only used to determine the initial radial distribution, not to set the upper bound. We substitute the Euclidean distances in global similarity terms Eq. (S10) by hyperbolic distances d^h_{ij} defined in Eq. (S2), and change Cartesian coordinate system to polar one for all the distance calculations. Then the gradient of total loss function with respect to polar coordinates would be:

$$\begin{aligned} \frac{\partial L_{total}}{\partial r_i} &= 4 \sum_j [(p_{ij} - q_{ij}) \cdot d^e_{ij} \cdot \frac{\partial d^e_{ij}}{\partial r_i} (1 + (d^e_{ij})^2)^{-1} \\ &\quad - \lambda(\hat{p}_{ij} - \hat{q}_{ij}) \cdot d^h_{ij} \cdot \frac{\partial d^h_{ij}}{\partial r_i} (1 + (d^h_{ij})^2)^{-1}] \\ \frac{\partial L_{total}}{\partial \boldsymbol{\theta}_i} &= 4 \sum_j [(p_{ij} - q_{ij}) \cdot d^e_{ij} \cdot \frac{\partial d^e_{ij}}{\partial \boldsymbol{\theta}_i} (1 + (d^e_{ij})^2)^{-1} \\ &\quad - \lambda(\hat{p}_{ij} - \hat{q}_{ij}) \cdot d^h_{ij} \cdot \frac{\partial d^h_{ij}}{\partial \boldsymbol{\theta}_i} (1 + (d^h_{ij})^2)^{-1}] \end{aligned} \tag{S14}$$

Where d^e_{ij} is the Euclidean pairwise distance in polar coordinates, d^h_{ij} is the hyperbolic pairwise distance obtained from Eq. (S2). p_{ij}, q_{ij}, $\hat{p}_{ij}$ and $\hat{q}_{ij}$ are defined by Eqs. (S5-S7) and Eq. (S10) with polar coordinates. When implementing the algorithm, we substitute the radial coordinates with their exponential transformation to avoid negative radii during the iteration:

$$r_{exp} = e^r \tag{S15}$$

The derivative of distances with respect to the new variable would be:

$$\frac{\partial}{\partial r_{exp}} = \frac{\partial}{\partial r} \cdot \frac{\partial r}{\partial r_{exp}} = \frac{1}{r_{exp}} \cdot \frac{\partial}{\partial r} \tag{S16}$$

When the iterations converge, we make logarithm transformation of r_{exp} to get the real radial coordinates.

1.4. Parameters in visualization algorithms

For Lukk et al. Lukk et al. (2010) data, we set $\lambda = 8$ and $R_{\text{model}} = 1$ in h-SNE. We select the result that best preserves data distances from 30 repeats. After obtaining the embedded points, we transform the points from native representation to Poincaré disk model by performing the transformation on radial coordinates:

$$r_{\text{Poincare}} = \tanh(\frac{r_{\text{native}}}{2}) \tag{S17}$$

In g-SNE, the parameter is: $\lambda = 20$. In PCA, we use the first two principal components for visualization. In UMAP, we screen a wide range of the combination of two key parameters: number of neighbors $\in \{5, 10, 20, 50, 100\}$ and minimal distance $\in \{0.001, 0.01, 0.1, 0.5, 0.8\}$, each of the 25 combinations was repeated 30 times. The optimal combination of parameters that leads to largest distance correlation of Shepard diagram is: number of neighbors = 100 and minimal distance = 0.5, and the corresponding result is shown in Fig. 6.For Moignard et al. data Moignard et al. (2015), the parameters for h-SNE are: $\lambda = 10$, $R_{\text{model}} = 1$. The root node index is 1800. When plotting Poincaré map, we directly use the embedding positions provided in Klimovskaia et al. Klimovskaia et al. (2020).

1.5. Evaluation of embedding

The Pearson correlation coefficient of Shepard diagram (embedding distances versus data distances) is used to measure the preservation of distances and global structure. For local clusters, we apply silhouette score Rousseeuw (1987) to our embedding results. Silhouette score measures the quality of data partitioning and clustering in graphical representation of objects, which in our case can be used to measure the consistency of the data configuration in 2D embeddings with the "ground truth" cluster labels. The higher score indicates better consistency with data labels. We consider all the three types of labels available – six hematopoietic properties, four malignancy properties and fifteen subtypes, and calculate the geometric mean of silhouette scores obtained by using these three labels:

$$s = \sqrt[3]{s_1 s_2 s_3} \tag{S18}$$

Where s_1, s_2, s_3 represent the silhouette scores by using the three types of labeling respectively. The mean score s is used to quantify the local structure preservation for the five visualization algorithms.

1.6. Data preprocessing and analysis

No pre-processing was done for the microarray dataset from human samples Lukk et al. (2010).

For scRNA-seq dataset from Han et al. (2018), we use Seurat packages Butler et al. (2018) to perform normalization, feature selection and scaling for the data, and to select the top 50 principal components for analyses. These results are reported in Figure 5.The results without pre-processing (and keeping all 1000 principle components) were qualitatively similar but had slightly reduced hyperbolic radii: $R_{\text{mouse brain}} = 1.62 \pm 0.02$, $R_{\text{mouse lung}} = 1.47 \pm 0.03$, $R_{\text{mouse kidney}} = 1.45 \pm 0.03$, and $R_{\text{mouse embryo}} = 0.98 \pm 0.02$ (compare with number in Fig. 5). These obervations are consistent with the observation that noise reduction done during pre-processing makes hyperbolic effects stronger and more apparent.

For mouse hematopoiesis data, we use the processed data from Klimovskaia et al. Klimovskaia et al. (2020). The Seurat analysis, violin plots and linear regression were performed using R version 3.6.2, the other analyses were performed using MATLAB R2017a.

2. Supplemental figures

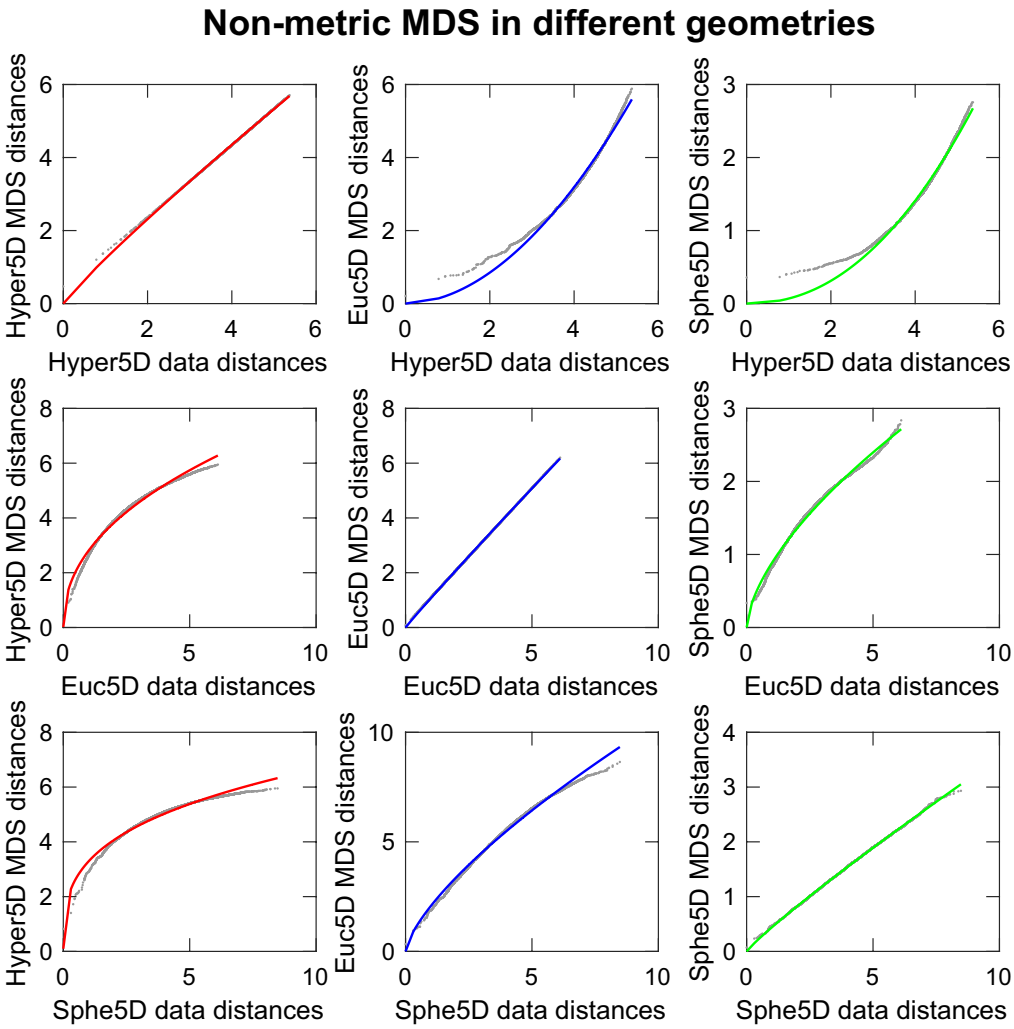


Figure S1: **Illustration of non-metric MDS embedding in different geometries. Related to Figure 2.** 100 synthetic points in 5D hyperbolic (top), Euclidean (middle) and spherical (bottom) space are embedded into 5D hyperbolic (left), Euclidean (middle) and spherical (right) space respectively. The hyperbolic radius is $R = 5$ for both data and model. The fitting methods are the same as in Figure 2 in the manuscript.

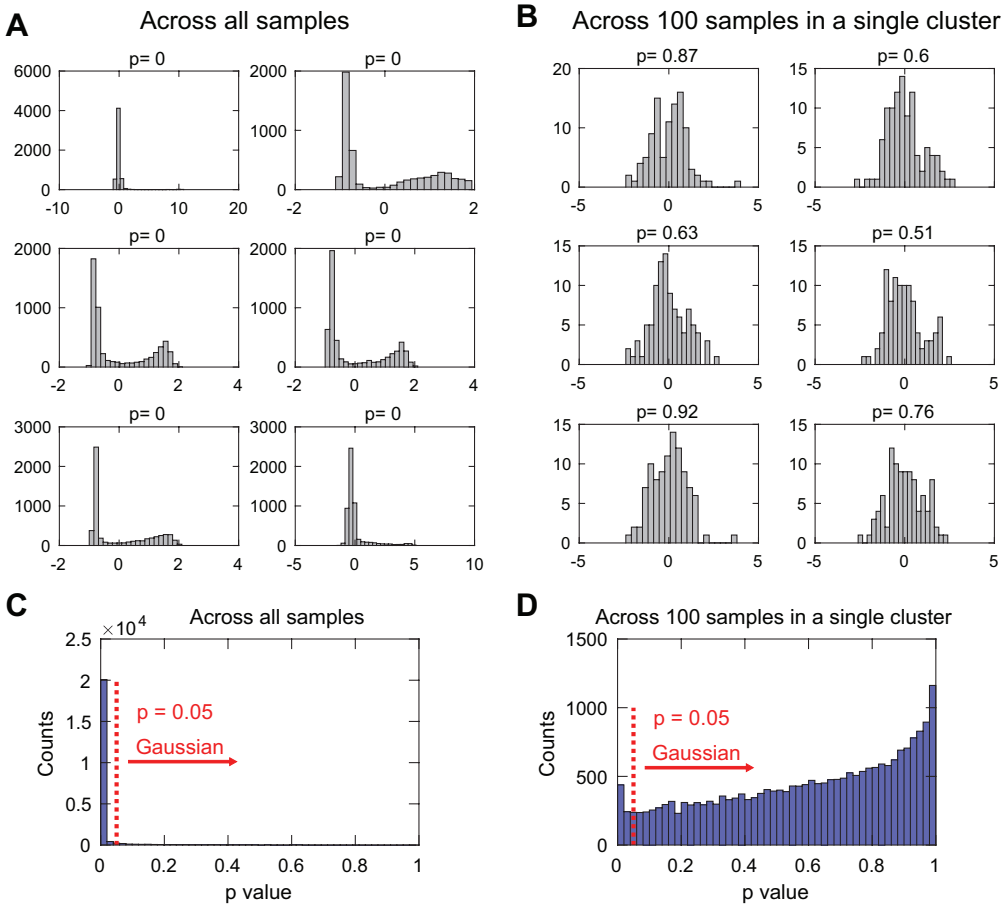


Figure S2: **Gaussianity of normalized gene expressions across the whole samples and from a single cluster. Related to Figure 3.** (A) Gene expression distributions of the six most non-Gaussian distributed probes across all the samples, p values were given by one-sample Kolmogorov-Smirnov test for Gaussianity, the null hypothesis is that the normalized distribution is standard normal distribution. (B) Gene expression distributions of the same six non-Gaussian probes as in (A) across 100 samples in one of the k-means ($k = 80$) cluster. (C) p value distributions of all the probes across the whole samples. (D) p value distributions of all the probes across 100 samples in one of the k-means cluster.

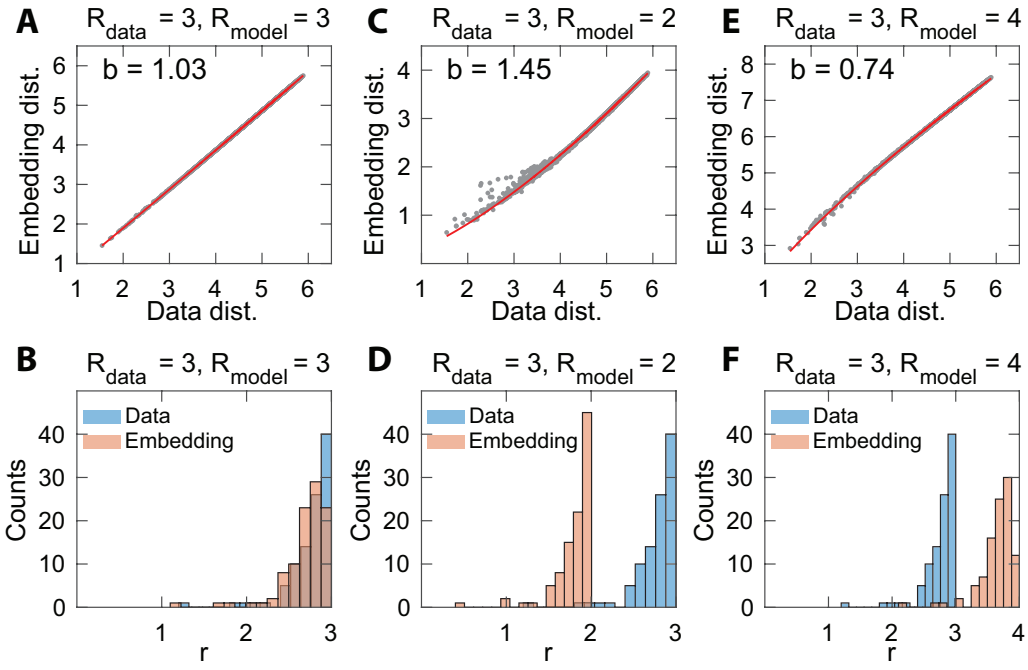


Figure S3: **HMDS embedding of hyperbolic data with different $R_{\mathbf{model}}$. Related to Figure 5.** 100 points are sampled in hyperbolic space with $D = 5$, $R_{\text{data}} = 3$. The embedding dimension is $D = 5$ in (A-F). The Shepard diagram convexity κ is shown in panel (A,C,E). (A) Shepard diagram of HMDS embedding of the samples to 5D hyperbolic space with $R_{\text{model}} = 3$. (B) Histogram of radial coordinates r of 100 sample points and model points after HMDS embedding with $R_{\text{model}} = 3$. (C-D) $R_{\text{model}} = 2$. (E-F) $R_{\text{model}} = 4$.

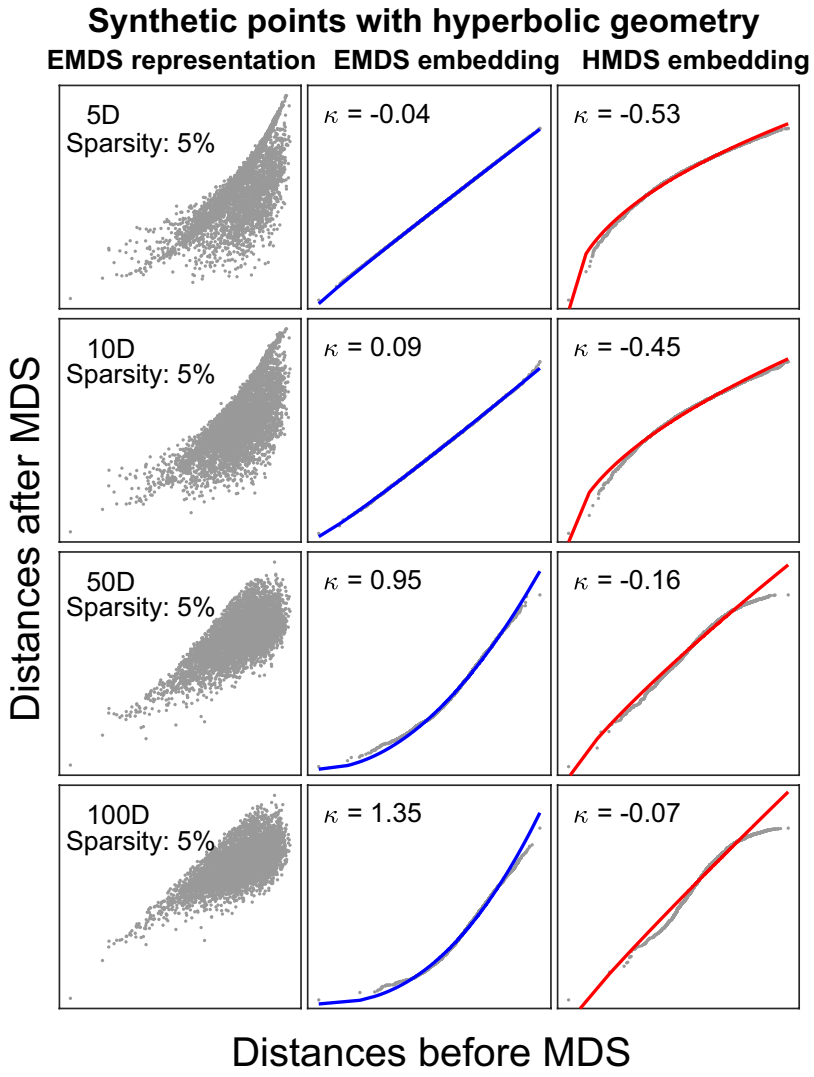


Figure S4: **Robustness to changes in data sparseness on geometry detection. Related to Figure 2 and Figure 5.** In Euclidean representation after EMDS, the coordinates of all the points are thresholded by fifth percentile of all the coordinate values to simulate the sparse RNA-seq values of cells. The left column shows the plots of thresholded embedding distances versus distances before MDS. The fitting plots and inserted convexity values in the rest columns have same meanings as in Figure 2 in the manuscript.

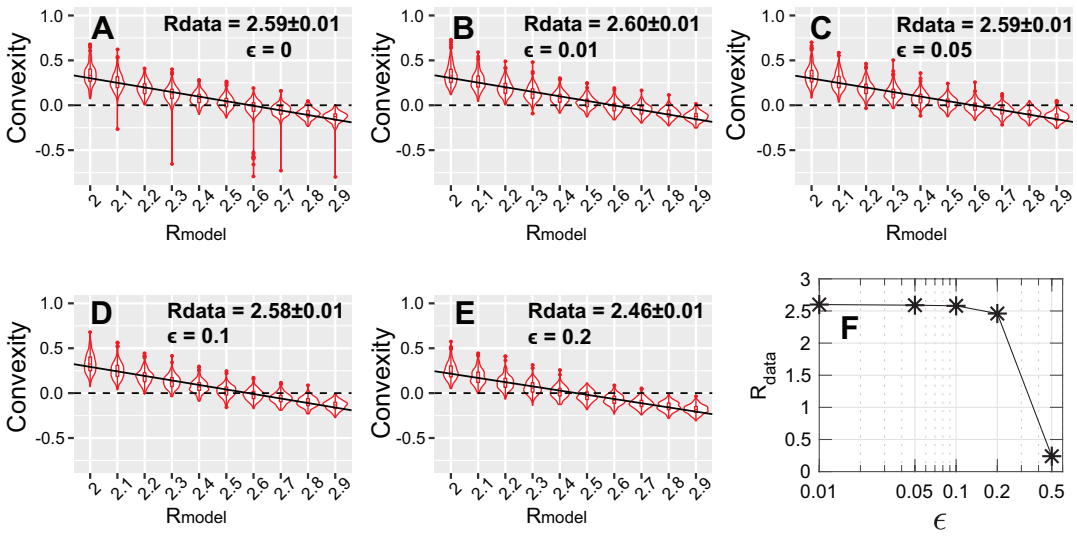


Figure S5: **Screening $R_{\mathbf{data}}$ of Lukk data with different magnitude of noise added by doing HMDS. Related to Figure 5.** The embedding dimension is $D = 5$. The noise ϵ is added as multiplicative Gaussian noise: $M_{noise} = M[1 + \epsilon \cdot N(0, 1)]$. (A-E) Fitting of R_{data} under different ϵ. (F) Plot R_{data} as the function of ϵ.

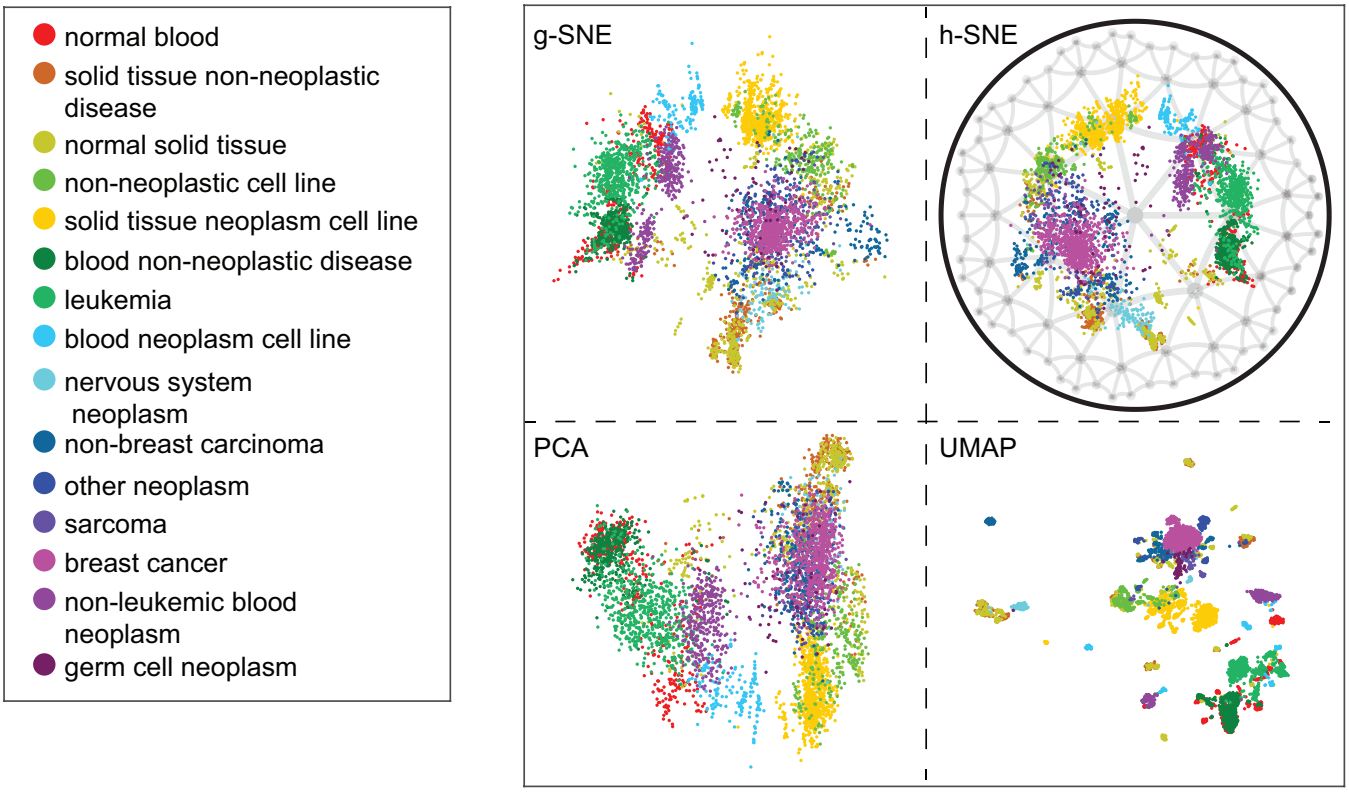


Figure S6: **Comparison of two-dimensional visualizations of human expression data using g-SNE, h-SNE, PCA and UMAP. Related to Figure 6.** The samples are colored according to the 15 tissue and disease types.

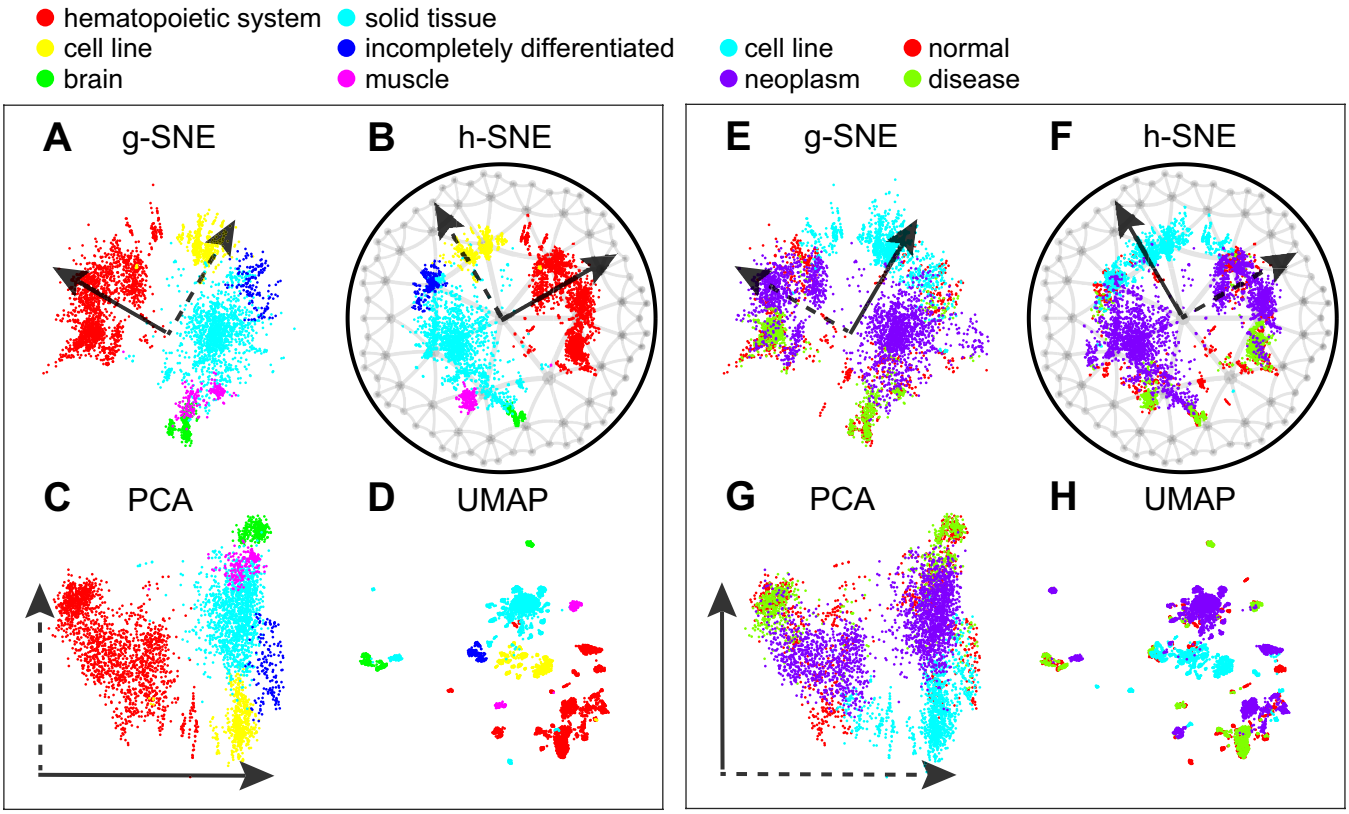


Figure S7: **Comparison of two-dimensional visualizations of human gene expression data in different algorithms. Related to Figure 6.** (A-D) g-SNE, h-SNE, PCA and UMAP embeddings of human samples classified by hematopoietic properties, these labels also represent the six major clusters identified by Lukk et al. The hematopoietic axes are shown in solid lines in g-SNE(A), h-SNE(B) and PCA(C). (E-H) g-SNE, h-SNE, PCA and UMAP embeddings of human samples classified by malignancy properties. The malignancy axes are shown in solid lines in g-SNE (E), h-SNE(F) and PCA(G). The six major clusters, the hematopoietic axis and the malignancy axis are hard to identify in UMAP.

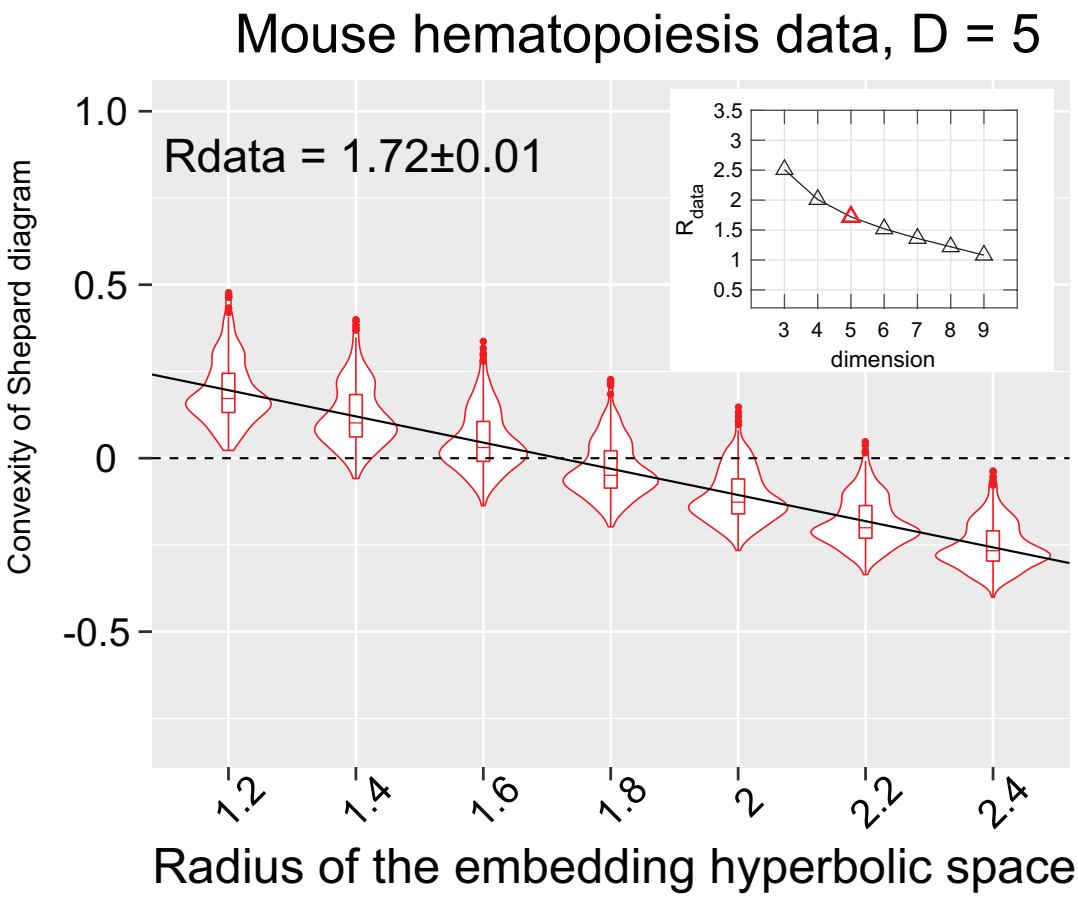


Figure S8: **Screening $R_{\mathbf{data}}$ of mice hematopoiesis data in Moignard et al. by doing HMDS. Related to Figure 7.** The inset shows the fitted R_{data} as the function of the embedding dimension.